The Economics of Tracking in Education

by Julian R. Betts,

UC San Diego and NBER

jbetts@ucsd.edu

This draft: September 15, 2009

First draft: August 21, 2009

This is a pre-publication version of Betts, Julian R., (2011), "The Economics of Tracking in Education", in Hanushek, Eric A., Stephen Machin and Ludger Woessmann (Eds.), **Handbook of the Economics Of Education**, Volume 3, Amsterdam: North Holland, pp. 341-381.

This paper was at a CESifo conference Economics in Munich, Sept. 3-5, 2009. I thank Ludger Woessman for helpful updates regarding the European literature on tracking. I thank Rick Hanushek, Ludger Woessmann, Derek Neal, David Figlio, Jeff Smith and Paul Ryan for helpful comments.

Introduction

Tracking refers to the tendency in many countries' public school systems to divide students by ability in some way. Students might be sorted into different classrooms within a school, or sorted into different schools. Definitions of tracking in the academic literature range from nothing more than ability grouping to more elaborate forms that divide students by academic achievement with the explicit intent of delivering a different curriculum, and using different pedagogical methods, for different groups of students. Countries differ widely on the degree to which they track students, and the age at which students begin to be tracked. Within-school ability grouping is common in the United States and Canada, while the practice of separating students into different flavors of schools at the secondary level is or was the case in many European countries for at least part of the last half century.

Proponents of tracking argue that it is economically efficient to group students by ability and perhaps by students' academic interests. By creating more homogeneous classrooms, in the case of tracking within a school, or more homogeneous schools, in the case of tracking across schools, in theory educators could tailor their pedagogical approaches for the given set of students. Separating students by initial achievement also opens up the possibility that school systems could tailor school resources to the given type of student. Examples of such resources include class size and teachers with certain types of qualifications. [1] At the secondary level, when students move beyond a focus on reading, writing and arithmetic to coursework that helps to prepare them either for postsecondary education or for more vocationally oriented jobs, tracking can also save resources by teaching each student exactly what he or she needs to know. In the case of countries that have tracked students into one of several tiers of secondary schools, one can imagine that the savings from not having to replicate every possible course sequence in any one school could be substantial.

Opponents of tracking argue that it condemns students placed into the lower tracks to lower educational attainment, and therefore lower earnings in their adult years. They fear that tracking aggravates economic inequality and perpetuates economic disadvantage across generations. Opponents have also criticized the academic data with which students are categorized into tracks, fearing that standardized test scores are untrustworthy. Student misclassifications seem particularly likely when tracking decisions are made in early grades. A final concern is that tracking changes the peer group of every student. The questions are whether those in the lower tracks suffer because of the reduced academic achievement of peers, and whether any such losses are counterbalanced by benefits to those in the higher tracks, whose peer group improves after tracking is implemented.

[1] As an example of why educators might want to customize the way they teach different groups of students, Finn and Achilles (1999) and Krueger (1999) show in the Tennessee class size reduction experiment that disadvantaged students tended to gain more from smaller class sizes. Betts, Zau and Rice (2003) report that smaller classes at the elementary level are particularly beneficial to English Learners, while Babcock and Betts (2009) present evidence that lower class size in elementary schools particularly benefits students who misbehave rather than the similar but distinct set of students who have low academic grades.

Tracking has generated heated political debate in many countries. The United Kingdom, which historically channeled students into three different types of schools based on a test administered to students when they were 11, began to move away from this system in the 1960's, based on fears that the system generated and perpetuated inequality. But due to political opposition to "de-tracking", remnants of the three-tier approach still exist four decades later in the United Kingdom. Sweden reformed its system in the 1950's, ending the assignment of students at age 12 or 13 to different educational paths. Other Scandinavian countries have implemented similar reforms. In the United States, schools widely practice within-school tracking. Many American researchers tend to dislike tracking on the grounds that it robs students from disadvantaged neighborhoods of the chance to achieve their educational potential. For instance, a widely cited book by Oakes (1985, 2005) provides a strong critique of tracking as it exists in America.

This paper will study what we know about the effects of tracking on overall student achievement and the distribution of student achievement, using studies from many different countries. One of the central problems in this diverse literature is the typically loose definitions of tracking that are available in existing data-sets. Similarly, researchers face several definitional problems when identifying the ability level of individual classes. A second central problem has been the endogenous selection of students into different tracks. A third problem is that school systems or countries may endogenously select a tracking or non-tracking approach. These problems have yet to be fully resolved.

Theoretical Foundations: Lessons for Various Empirical Approaches

To know what questions to pose about the economic consequences of tracking, we must begin with a clear idea of the factors that contribute to student learning, and how various forms of tracking might affect these basic factors. One can easily imagine an education production function that goes beyond the standard inputs of class size and teacher qualifications, and that in addition allows for multiple paths through which tracking might affect an individual student's achievement. The most obvious mechanism is that when a school adopts tracking within its classrooms, or a district adopts tracking across its schools, each student's peer group changes. Teaching style (pedagogy) and subject matter (curriculum) could also vary across classrooms when schools track, especially in secondary schools and in the European context, where in certain countries secondary students are streamed into entirely different types of schools. Examples of how pedagogical approaches might vary when tracking is used include variations in how the teacher divides her time, first, among students, and second, among whole-classroom, small group, and individual instruction.

We start with a production function for the test score for student i in country c, school s, with teacher r in grade g and year t, S_{icsrgt}. Assume that this test score depends on innate ability A_i (in a way that may vary with age), other personal traits X_{icgt}, a vector that implicitly contains an overall intercept, some key family characteristic F_{icgt}, a vector of teacher qualifications $QUAL_{icsrgt}$, teacher effort $QEFF_{icsrgt}$, which we index with i because the teacher may focus her efforts in a way that helps some students in the class more than others, class size $CLASS_{icsrgt}$, measures of the initial achievement and perhaps other

traits of classroom peers, captured by $PEER_{icsrgt}$, a vector of variables indicating the curriculum being taught, captured by $CURR_{icsrgt}$, and a vector of variables indicating the pedagogical approach taken by the student's teacher PED_{icsrgt}. School level characteristics $SCHOOL_{icsgt}$ could matter. Finally, parents may purchase educational activities or materials for their child after school, $PRIV_{icgt}$. For example, many Japanese parents pay to send their children to after-school tutoring in private schools known as *juku*. It is likely that the entire past school history of the student affects his or her current score, but we omit these lags for the sake of simplicity:

$$(1)\quad S_{icsrgt} = f\left(\begin{array}{l} A_i, X_{icgt}, F_{icgt}, QUAL_{icsrgt}, QEFF_{icsrgt}, CLASS_{icsrgt}, \\ PEER_{icsrgt}, CURR_{icsrgt}, PED_{icsrgt}, SCHOOL_{icsgt}, PRIV_{icgt} \end{array}\right)$$

Suppose that some schools use tracking. Then all of the determinants of achievement in (1) apart from A_i, X_{icgt} and F_{icgt} are likely to be endogenous functions of both the overall use of tracking and the specific track to which a student i is assigned. With tracking, school administrators are able to alter all of the classroom and school characteristics. For instance, the use of tracking is likely to influence the qualifications of student i's teacher, and class size, because administrators now have the ability to tailor the resources devoted to students at different achievement levels. If teachers react differently to having more homogeneous groups of students in tracked schools or to higher or lower ability groups within tracked schools, teacher effort $QEFF_{icsrgt}$ will also depend on the use of tracking and perhaps student i's specific track. Peers, curriculum and pedagogical approaches adopted by the student's teacher could also depend upon the use of tracking. In countries such as Germany and Italy in which students are tracked into different types of secondary schools along the vocational:college preparatory continuum, the vectors of school characteristics $SCHOOL_{icsgt}$ will also depend on tracking.

More subtly, if parents make decisions on how much to spend on private tutoring, $PRIV_{icgt}$, in response to the quality of schooling that is being provided for their children, this spending too could change once tracking was implemented. For instance, in a system without tracking, more affluent parents may spend considerably on private tutoring because they want to find ways to create a separating equilibrium in which their children obtain the top grades in school and gain either the best postgraduation jobs or admission to university. Such spending might fall once tracking were introduced because parents viewed placement of their children into the upper track as a substitute for private tutoring. But less affluent parents may do the opposite if they perceived that once tracking was instituted, their children were likely to be placed in the lower tracks. Because affluent parents have greater financial resources than do less affluent parents, it is likely that the institution of tracking would lower overall parental spending on private educational resources.

This insight may hold implications for studies that use variations in tracking policy over time. If affluent parents seek a separating equilibrium that benefits their children, then they will increase private tutoring expenditures if tracking is ended. Presumably, then, this would bias down the estimated effects of tracking on both the equality of student outcomes and the overall level of student outcomes.

There are many complications we could add to the model. For example, one of the main concerns expressed in the empirical literature on tracking is that tracking in one grade affects student outcomes and the sorts of courses and teachers to which students have access in later grades. Thus each of the determinants of test scores in grade g in (1) depend on tracking not only in that grade but in prior grades. The above production function is also ripe with possibilities for interactions between the various inputs on the right hand side, and other types of non-linearities. [2]

What decisions must a researcher make before attempting to estimate a version of the potentially complex production function implied by (1)? Suppose that a longitudinal student-level data-set becomes available. Even if the data contained detailed information on peers, pedagogical approach, and so on, the endogeneity problems would be severe. In the student-level studies that have obtained information on what ability group a student has been assigned, researchers typically have estimated a reduced-form that has not attempted to include controls for teachers' pedagogical methods, peer achievement and other endogenous factors, but instead has included either a simple control for whether tracking is used or an interaction between the use of tracking and the student's ability level, and some controls for overall school resources.

Similarly, the many studies that have used international variation in the use of tracking typically omit the many endogenous explanatory variables in (1).

Potential for International and Across-School Variations in the Meaning of Tracking, and for Mismeasurement

While this tendency to estimate reduced-forms makes sense, it does raise serious questions about what we mean by tracking. Especially in the many international studies in this literature, we are estimating an average treatment effect across what might be really quite different types of tracking between one country and another. For example, one country that tracks may stream students into different schools, with different types of teachers and other resources, and intentionally different curricula. In another country that uses tracking, tracking could be within-school, and it might give largely the same curriculum to different students, with similarly trained teachers, but use different pedagogical approaches for students in the various tracks.

Such differences in the meaning of tracking seem quite possible. A characterization of many European approaches to tracking is that it separates students into two or three different types of schools, each of which has quite different curricular focus. American and Canadian schools are typified by informal tracking that rarely involves sending students to different types of schools. This will become an extremely important point to bear in mind when we consider studies that make international comparisons.

[2] Non-linearities raise concerns that aggregated analyses that average over large numbers of schools could be subject to aggregation biases. See Theil (1954) for an early examination of the issue of non-linearities and Hanushek, Rivkin and Taylor (1996) for a study specific to education production functions.

For instance, Brunello and Checchi (2007), similarly to other authors who compare tracking across countries, report that the age at first tracking in the United States is 18, which is true only if one ignores the very strong within-school ability grouping and streaming that occurs as early as middle school. See e.g. Oakes (1985, 2005.)

Endogeneity of Tracking

For both student-level within-country and international studies, even if we drop endogenous mediating variables such as teacher qualifications and teacher effort, the endogeneity of tracking itself remains an issue. Why do some schools within a country track, and why do some countries but not others track? The factors determining whether public schools track are likely to be quite complex, and in many cases these factors could affect achievement of students through other paths than tracking itself. For instance, the degree of competition from private schools may affect the probability that public schools within a region or country adopt tracking, as hypothesized by Epple, Newlon and Romano (2002). Yet competition from private schools is likely to affect public school achievement through diverse channels in addition to whether public schools track. Societies that are more conservative in the sense of having less egalitarian social policies are more likely to use tracking to benefit the children of more affluent families. Societies that are more socially divided may seek out tracking as a way of separating students from different types of backgrounds. Racial mix, religious mix and immigrant/native mixes could also influence the use of tracking.

Figlio and Page (2002) present results indicating a positive association between changes in American schools' use of tracking and changes in the socioeconomic status of students. Their interpretation is that more affluent students seek out schools that track, although of course the causation could run in the opposite direction. They also find that county voting patterns, state graduation requirements, and the amount of public school competition are correlated with whether a given high school tracks.

For within-country studies, the endogeneity of tracking can become worse than in the case of international studies. First, families decide where to live, and so may opt into or out of a school with tracking through their locational choice. Second, if a data-set contains specific information on the track into which a student has been placed, it may be dangerous to condition on this information because this placement is itself endogenous.

On the other hand, across-country studies may suffer from greater omitted variable bias than studies using variation within a country, because of greater unobserved heterogeneity across countries than across areas within a country.

A Benchmark Data Generation Process

With these data concerns noted, it is useful to compare the production function in (1) with the sort of model that is estimated in the majority of studies that have used variations in the use of tracking across geographic areas. The example below assumes that the geographic unit at which tracking is

observed is countries, but the estimating equation would carry over to a case in which tracking varied by county or other geographic unit within a country.

We assume that we have no data on the range of endogenous variables in (1) that schools might alter if they were able to use tracking. But we assume that some measure of overall school resources is available, without which omitted variable bias could be quite severe. We also re-introduce the idea alluded to briefly above that the entire history of a student's experience with tracking should affect his or her current achievement.

Consider the following data generation process, in which we observe the test score for student i in country c, in grade g and year t. Assume that this test score depends on personal traits X_{icgt}, a vector that implicitly contains an overall intercept, some key family characteristic F_{icgt}, some measure of the average school resources, such as spending per pupil, that student i experienced from kindergarten up to grade g, Q_{icgt}, and indicators for whether the student in each year and grade was in a tracked school (or lived in a country in which students were typically tracked in that year and grade), captured by indicator variables T_{icgt}, $T_{ic,g-1,t-1}$ and so on. We assume that the data generation process (d.g.p.) might be given by:

$$
\begin{aligned}
S_{icgt} = {} & X_{icgt}\Phi + F_{icgt}\Pi + Q_{icgt}\gamma + T_{icgt}\lambda_g + T_{ic,g-1,t-1}\lambda_{g-1} + \mathrm{L} \; + T_{ic,1,t-g+1}\lambda_1 + \\
(2) \qquad & T_{icgt}F_{icgt}\rho_g + T_{ic,g-1,t-1}F_{ic,g-1,t-1}\rho_{g-1} + \mathrm{L} \; + T_{ic,1,t-g+1}F_{ic,1,t-g+1}\rho_1 + \\
& \left(\alpha_i + \alpha_c + \alpha_g + \alpha_t + \beta_{cg} + \beta_{ct} + \beta_{gt} + \delta_{cgt} + \varepsilon_{icgt}\right)
\end{aligned}
$$

The λ_k terms capture the average effects of having been tracked in grade k. [3] The model includes interactions between the tracking dummies T and the family characteristic F. The ρ_k terms in (2) capture the differential effects of tracking on different socioeconomic groups, and thus are of interest to those focusing on whether tracking generates inequality in education outcomes. Some researchers exclude these terms in some of their models, so that in these simplified models the λ_k terms provide a measure of the overall efficiency effect of tracking, holding constant school resources Q_{icgt}.

A comparison of (1) and (2) suggests that we may be missing some important explanatory variables, but that we have written the reduced form of a plausible data generation process. Researchers who typically estimate a variant of (2) and who exclude endogenous regressors such as teacher qualifications and curriculum will thus also have sidestepped some severe issues of endogeneity bias. [4]

[3] Note that for simplicity we assume that the effects of being tracked in a given grade are permanent in the sense that they have the same impact on test scores in grade k and all later grades. If we allowed for deprecation of the effects slight complications which are well understood would be added. One could also imagine non-linear effects on a student of being tracked for multiple years. These issues have yet to be studied empirically.

[4] For simplicity we have conditioned upon a student's past experience with tracking, but not upon past school resources. Researchers lack the detailed educational histories needed to do this, but sometimes condition upon a

Tracking and School Resources

Before studying the effects of tracking on achievement, it is helpful to characterize how resources vary across tracks, and how countries themselves might vary in how they implement tracking systems.

Betts and Shkolnik (2000a) provide the first characterization using a large nationally representative U.S. dataset of the school resources assigned to students in the five ability groups. They find that, relative to teachers of the top classes, teachers of the lowest ability classes tend to have less experience (12.3 years versus 15.0 years for the top classes) and to be less likely to hold a Master's degree (50% versus 69%). But conversely, class sizes are smaller for the bottom ability group (at about 19 students compared to about 26 students in the top ability group). It is possible that these smaller class sizes allow teachers to spend more time on individualized instruction. Evidence for such a correlation is provided in a separate study by Betts and Shkolnik (1999).

Betts and Shkolnik (2000a) also find some subtle differences in these patterns between schools at which principals claim ability grouping is used and schools at which principals claim ability grouping is not used. Most importantly, they find that only in schools with formal grouping do class size and teacher experience fall substantially for the bottom-ability classes.

Rees, Brewer and Argys (2000) replicate these results using a separate nationally representative U.S. data-set. They report that grade 10 classes in history, math, science and English tend to be smaller for below-average ability classes (by about 3 or 4 students), and that for math (only) students in low-ability classes are less likely to have a teacher with a Master's degree. They also report very small differences in teacher experience across ability groups, but with a very weak pattern in which low-ability classes receive teachers with lower experience.

In contrast, Brunello and Checchi (2007, page 795) document large differences in pupil-teacher ratios experienced by students in different tracks in upper secondary grades in various European countries. They write that the largest differences in pupil-teacher ratio are in Germany, with 11.89 students per teacher in the general track and 21.25 students per teacher in the vocational track. Corresponding figures are, for Austria, 9.05 versus 14.51, for France, 6.75 and 14.67, and for Italy, 11.17 and 11.94. Conversely, though, they report calculations on total expenditures per student in vocational and academic tracks in Austria and find that the former is generally higher, which could perhaps arise due to the costs of on-the-job training components in the vocational tracks.

These pupil-teacher ratios are not the same thing as class size – actual class sizes will be larger because teachers typically have preparation time outside the classroom each day. But still, these figures show the opposite pattern to American schools in that class sizes appear to be larger in the vocational track in these European schools.

---

lagged score as a proxy for these past experiences. This imposes strong restrictions. An even more restrictive version of such a model models gains in achievement.

Perhaps in Europe it is politically more feasible to implement bigger resource differences by track because at the secondary level many European countries send different ability groups to different types of schools. Funding two or three types of schools differentially may be less visible (and objectionable) than would be the same resource differences implemented along different corridors of the same school, as would be required in the U.S style of within-school ability grouping.

In the following sections that evaluate the effects of tracking, it will be helpful to bear in mind the contrasts between the American and European versions of tracking. First, as mentioned in the previous section, American high schools tend to offer various tracks under the same roof, while the European approach more typically houses vocational and academic tracks under separate roofs. Second, as shown here, the differences in classroom resources in American high schools, as reported in two papers, are not large, and no ability group receives more of all resource types. But in four European countries pupil-teacher ratios are unequivocally larger in vocational schools.

The overall sense that emerges, in which European versions of tracking are more dramatic, will prove useful in reconciling differences between the American and international literatures on the effects of tracking on student achievement.

Empirical Approaches to Estimating the Effects of Tracking

This section outlines the main approaches used thus far to estimate the effects of tracking. Each method is outlined, along with the main technical challenges in each case, and a review of findings.

Traditional across- and within-School Variation

Many non-experimental papers compare students in different tracks within a school or across schools. These regression-based studies typically use a sample of schools within a country or a smaller geographic area. In relation to the formulation in (2) most of the terms in the error term are not accounted for by adding fixed effects, apart from dummies for grade level α_g. With one exception, this literature has taken the decision by schools to use tracking as exogenous. Typically but not always, researchers have assumed that the endogeneity of the student's track can be controlled for by including a sufficiently rich set of covariates.

Early Work

Slavin (1987, 1990) assesses the large early literature, almost all of which was written by social scientists outside of economics. Researchers have typically estimated two types of equations, which aim to answer either the question of whether tracking affects efficiency of schools or the question of whether tracking contributes to inequality in outcomes. To answer the first question, researchers have compared overall achievement at schools with and without tracking, ceteris paribus. To study the

second question, a single explanatory variable for tracking is typically replaced with the ability level of the class to which the student was assigned.

Slavin (1990) reviews 14 regression-based student-level studies of tracking in secondary schools. The studies he includes focus on American and to a lesser extent British schools. On the question of efficiency, Slavin shows that effect sizes in studies that compare tracking to non-tracking schools vary but are typically close to zero. Somewhat more surprisingly, on the question of whether those in high, medium and low ability groups in schools with tracking perform differently from heterogeneously grouped students, he again finds no consistent patterns. Most of these studies involved anywhere from 1 to 28 schools. Thus, the often insignificant results could result from the low statistical power of the smaller studies. The most important exception to the characterization of the early literature as being small scale is Kerckhoff (1986), which studies a large sample of secondary school students in the United Kingdom. This paper concludes that there are small positive effects on average achievement, and that students in the high-ability classes had significantly higher test scores than students in the low-ability classes, after controlling for initial achievement.

Slavin (1987) provides a review of early work on tracking at the elementary school level. He reports that 13 correlational or regression-based studies typically find no overall effect of "comprehensive" (school-wide) ability grouping on achievement, and mixed evidence on whether students in higher-ability classes gain more than students in lower-ability classes. Again, many of these early studies use small samples, with eight using samples of under 1000 students, and most studying either a handful of schools or, in one case, two districts, one with and one without tracking. The two largest studies, one conducted in England and Wales and the other in New York City, reported no or slightly negative overall effects of ability grouping, and no evidence that high ability groups perform better when they are in high-ability classes.

Slavin (1987) presents far more favorable evidence for two specific types of ability grouping in elementary schools. The first, known as the Joplin plan, allows for regrouping of students across grades for reading, with the grouping based on initial reading prowess. Eleven of 14 studies, including two experimental studies, suggested that this approach led to increases in average reading achievement, with the median effect size 0.45. Second Slavin summarizes 8 studies of within-class ability grouping, including five experimental studies. All but one study, which examines math, reading and spelling, focus on mathematics. All studies suggested that within-class ability grouping led to an increase in average achievement. The studies differed on whether students in high- or low-ability classes gained more from this type of grouping.

Slavin's literature reviews have been quite influential, and were seen as a rebuttal to the claims by Oakes (1985, and also 2000), which were based on observation of individual schools, that in the United States tracking hurts low-performing students.

However, concerns about the papers reviewed by Slavin (1987,1990) include not only small sample sizes but the very limited geographical range of the individual American studies, and limited attention to the endogeneity of group placement.

Newer American Research that Uses Nationally Representative Samples

More recently, in the American literature, a number of papers have used far larger longitudinal data-sets than were available in the early studies. Moreover, these data-sets were also nationally representative. Hoffer (1992) studies math and science achievement of middle school students, and makes the important step of attempting to control for selectivity bias in the track to which each student is assigned, using propensity score methods. He finds that students placed in the upper group in schools that use ability grouping outperformed otherwise similar students in schools without tracking, while those placed in the low-ability groups underperformed. When he examines the overall effect on student achievement, tracking has no statistically significant effect. Thus, it would appear that tracking increases inequality without boosting efficiency.

Gamoran and Mare (1989) similarly use a nationally representative sample and, even after attempting to control for selectivity, find that in high schools tracking tends to increase inequality.

A paper by Argys, Rees and Brewer (1996) also yields results different from the general finding in Slavin's reviews of no or small effects. This paper uses Heckman selectivity corrections to control for assignment to track. In models of grade 10 math achievement, the authors report that tracking dramatically increases inequality, boosting scores of students in medium- and high-ability classes while lowering the achievement of students in the low-ability track. They find that tracking boosts average achievement slightly, by about 2%.

Betts and Shkolnik (2000a,b) introduce other approaches to studying tracking, again using a nationally representative set of secondary schools in the U.S., and raise concerns about bias in the approaches used by Argys, Rees and Brewer (1996) and Hoffer (1992). Their first innovation is to reduce the selectivity bias problem inherent in comparing students in high-ability classes in schools with tracking to all students in heterogeneously grouped classrooms. They use reports by the principal on whether the school uses ability grouping, combined with reports that are available for *all* schools on the ability level of students in classrooms. This allows them to compare student outcomes in classes of a given ability level in tracking versus non-tracking schools. They find little difference between outcomes for students in a class of given ability level between schools in which the principal reports the use of tracking and schools in which principals claim that tracking is not used.

Rees, Brewer and Argys (2000) reinterpret the approach used in Betts and Shkolnik (2000a), arguing that if teachers in all schools are prepared to rate their classes' average ability level, then it must be the case that all schools use ability grouping. Thus, they argue, Betts and Shkolnik (2000a) are not comparing ability grouping to the absence of ability grouping, but instead are comparing formal ability grouping to informal ability grouping.

This is a reasonable interpretation of the approach adopted by Betts and Shkolnik (2000a), but the larger point in the Betts and Shkolnik (2000a) paper is that the meaningful effects on inequality reported by Hoffer (1992) and Argys, Rees and Brewer (1996) reflect inadequate controls for endogenous placement of students into tracks.

Betts and Shkolnik marshall several pieces of evidence in this regard. Betts and Shkolnik (2000a) replicate the comparisons made by the aforementioned authors, by comparing achievement gains of those in a given track in schools with ability grouping to a control group consisting of *all* students at schools where ability grouping was not used. Like Hoffer (1992), who uses a subsample of the same data-set as Betts and Shkolnik, they find large positive and negative effects of being in a high-ability and low-ability class, respectively. But Betts and Shkolnik (2000a) express skepticism about the size of the effects. Figure 1 shows the 25^{th}, 50^{th} and 75^{th} percentiles of actual test scores in math by grade in their nationally representative sample, and superimposes on this (with dotted lines) the predicted test scores by grade of two hypothetical identical students who performed at the median level at the start of grade 7, but who were randomly assigned to the top and bottom ability groups in ensuing years. Betts and Shkolnik (2000a) conclude that this rapid divergence between identical students predicted by standard models is far too big to be realistic. They reason that in their national sample the variance of test scores changes little across grades, and yet the vast majority of students (73%) in that sample are in schools that use ability grouping. They conclude that standard models do a poor job of controlling for unobserved differences among students.

The second piece of evidence that some of the earlier U.S. papers overstated the effect of tracking on inequality comes from models estimated by Betts and Shkolnik (2000a) that use Heckman selectivity corrections, propensity score, and instrumental variables approaches to control for endogenous group placement. They use as instruments students' lagged test scores, divided by average lagged scores in the students' school and grade, as well as school demographic variables, as predictors of the ability group to which the student is assigned. Once they use these explanatory variables and any of these three methods, the differential effects of ability grouping largely disappear. This does not appear to be due to weak explanatory power of the added instruments, which have considerable explanatory power in the first stage of the instrumental variables procedure.

Betts and Shkolnik (2000b) present other evidence that the approaches used by Hoffer (1992) and Argys, Rees and Brewer (1996) almost surely do not control adequately for endogenous group placement. They point out a thoughtful robustness test used by Hoffer (1992): instead of conditioning just on test scores from the previous year, he also conditions on test scores from two years earlier. Once he does this, the predicted gap between those in the high and low ability groups drops by about one half for math and one third for science. More to the point, although in his original model all three of his class ability levels were statistically significant, the only ability group coefficient that continues to be significant at the 5% level after adding the twice lagged test score is that for the low ability group. This provides clear evidence that the norm in these papers of controlling for demographics and lagged achievement does a poor job of controlling for omitted ability bias: teachers assign students to ability groups based on a fuller knowledge of the student's actual achievement and motivation than can be gleaned by researchers who typically must rely upon a single noisy test score.

In the case of Argys, Rees, and Brewer (1996), Betts and Shkolnik (2000b) question this paper's implementation of Heckman selectivity corrections. First, none of the Inverse Mills terms is significant, indicating either that students are assigned randomly to ability groups (conditional on observables), which seems highly unlikely, or that the first-stage model of track placement lacks explanatory power.

Second, they point out that the signs of the Inverse Mills terms are incorrect in that they suggest that students in the high ability track were negatively selected, and that students in the low ability track were positively selected.

In a similar vein, Figlio and Page (2002) use the same data as Argys, Rees, and Brewer (1996), and replicate their results fairly closely, but then question whether it makes sense to treat students' track placements as exogenous. Instead of attempting Heckman selectivity corrections, they replace the ability group variable with indicators for whether initial test scores of the individual student were in the bottom, middle or top third of the grade 8 distribution. They find no effect of being in a tracked school for any of these three groups.

In a first in the literature, Figlio and Page (2002) use instrumental variables for the existence of tracking at the school, using county-level instruments. They report fairly good first-stage fit and if anything, their results support positive effects of tracking for students in the bottom of the initial test-score distribution and zero effects for students in the top two-thirds of the initial distribution.

Takeaway Lessons

Betts and Shkolnik (2000b) close with six observations for researchers who decide to undertake future non-experimental work based on student/school level data on the effects of tracking. In brief, and with minor amendments, these are:

1) A student's classroom ability group is likely correlated with unobserved ability and motivation in regression-based studies. Therefore, unless adequate precautions are taken, the effects of tracking on inequality are almost surely overstated.

2) Informal ability grouping appears to be extremely common in American public schools, making it difficult to find a true "ungrouped" school.

3) Some studies such as Argys, Rees and Brewer (1996) use as the comparison group classes that the teacher has labeled as "heterogeneous" ability. This is a very vague term that could mean different things to different teachers, and care must be used when using such definitions. (By the same reasoning, surveys that ask teachers to label the ability level of their class may not be particularly reliable, a point made by Rees, Brewer and Argys (2000).)

4) Few U.S. studies at the secondary level make a clear distinction simple ability grouping and ability grouping that is combined with differences in curriculum or pedagogy. The effects of the two types of ability grouping could be quite different.

5) We need to know more about whether and how secondary schools use ability grouping as an opportunity to tailor class size, teacher qualifications and other inputs to the needs of students.

6) We know little about how schools group by ability within classrooms, especially at the secondary level.

Overall, this literature does not provide compelling evidence on either question – the overall effect of tracking on average achievement, or the effect of tracking on the distribution of achievement. Oakes (1985, 2000) writes powerfully about the disadvantages faced by students who are placed in lower tracks in American schools, and states that classroom observations indicate that students in low tracks spend less time on task in the classroom. Yet only a few studies in the quantitative literature find strong differential outcomes on standardized tests, and some of the papers that do find strong signs that tracking increases inequality appear to suffer from considerable omitted variable bias and endogeneity bias. Slavin (1990) notes that Oakes' observation that low-track students tend to spend less time on task, although legitimate, does not provide evidence that these differences are exacerbated by tracking. Rather, he writes: "Is this due to the poor behavioral models and low expectations in the low-track classes, or would low achievers be more off-task than high achievers in any grouping arrangement?". (Slavin, 1990, p. 474)

Again and again, the problem that crops up in the regression-based literature is the extreme difficulty social scientists face when attempting to estimate the counterfactual without tracking. Indeed, attempts by Betts and Shkolknik (2000a,b) and Figlio and Page (2002) to take into account the endogeneity of group placement suggests that tracking does not aggravate inequality in academic achievement, even though simpler models suggest it does.

A final observation is that with one exception this literature treats the use of tracking by a school as a decision that is exogenous with respect to student achievement. Any unobserved factors that are correlated with the use of tracking, and which are related to achievement, will bias the results of the papers reviewed above. Figlio and Page (2002) provide evidence that treating the existence of tracking at a school as endogenous can reverse earlier findings by Argys, Rees and Brewer (1996) that tracking aggravates inequality.

Approaches that Geographically Aggregate Schools

A large parallel literature avoids the difficult questions related to which individual schools track, who attends them, and the tracks chosen by individual students, and instead geographically aggregates schools, thus facilitating comparisons between one region and another or one country or another.

Geographic:Time Difference in Differences: Natural Experiments Related to Policy Reforms

Numerous countries or regions within countries have changed educational policies related to tracking. If it can be argued credibly that the change in tracking is exogenous with respect to student learning, and not correlated with any other (unmeasured) reforms to education policy, then one may have a plausible natural experiment for identifying the effects of the reform to tracking. Most of the studies to date have examined the phase-in or phase-out of tracking across regions of a single country.

This geographical identification approach holds some clear advantages over studies that make school-by-school comparisons. The international and sub-national papers, through their use of geographic aggregation, mitigate concerns about the endogeneity of the decision to track at the school level, the endogenous assignment of individuals to specific tracks, and the endogenous choice by families of where to live. Obviously, the studies of regions within a country will reduce biases due to endogenous residential choice, but not as much as international studies, because families are free to move among regions within a country.

The difference in difference approach and its strengths and weaknesses are well known. Referring to our d.g.p. in (2), the typical paper in this literature explicitly includes dummy variables for geographic unit and time, removing the α_c and α_t from the error term. As always, the success of this approach depends on the assumption that trends between the control and treatment units (areas without and with tracking) must be the same, apart from differences directly resulting from tracking and other observables. Any correlation between the tracking variables and the error components β_{ct} and δ_{cgt} in (2) will lead to bias. We will therefore pay attention to whether any other changes within geographic units may have occurred at the same time as the institution (or abolition) of tracking.

Meghir and Palme (2005) study student outcomes in the context of major policy changes in Sweden. Until roughly 1950, all students attended compulsory elementary schools up to sixth grade, at which point students with the best grades enrolled in junior secondary schools, which had a strong academic focus, with these students later articulating into upper secondary schools and ultimately, postsecondary education. Students in grade 6 who had lower grades were required to attend more basic compulsory schools, for either one or two additional years, and had the opportunity to attend vocational schools after that. Between 1949 and 1962, Sweden experimented with a new approach that implemented several reforms at the same time. Most relevant for our purposes, the reform ended placement into academic versus non-academic tracks at the end of grade 6, and also introduced a national curriculum for all secondary students.

This reform did not require that all schools teach exactly the same material. Indeed, a three-level secondary school system was created, with academic, more basic academic, and vocational paths available at the student's discretion. Individual schools typically housed all three of these programs.

Notably, the reform also went further, increasing the minimum number of years of schooling required from seven or eight years (depending on the region) to nine years. Further, this increase in the school attendance requirement was buttressed by a financial stipend to families to make up for the lost labor-market earnings of adolescent family members who would have otherwise entered the labor market had the school attendance law not changed. As the authors point out it is impossible to know for sure which elements of the reform caused observed changes to students' earnings years after leaving school.

Meghir and Palme adopt a difference- in-differences approach that takes advantage of the fact that the reforms were phased in. They compare outcomes for two birth cohorts, the older of which was more likely to have experienced the old system. In addition they exploit variation across municipalities.

(A national board selected which municipalities would implement the reform in a given year.) The key variable in their outcome model is an indicator for whether the person was in a cohort and lived in a municipality (in grade 6) that was subject to the new system. Outcomes include earnings and various measures of educational attainment.

They find that on average the reform was associated with an increase of 0.3 years of schooling completed. Students whose fathers had low education accounted for all of this gain. Most of the gain derives from the increase in the required years of school attendance, but attendance beyond the compulsory level also rose by 2.6 percentage points. Clearly, the increase in years of compulsory attendance induced much if not all of the increase in enrollment and attainment. It is not possible to infer whether the end of tracking contributed to these increases.

Meghir and Palme report a 1.4% increase in earnings that is associated with the reform package, but the change is not significant at conventional levels. Nonetheless, in the new regime earnings increased 3.4% for workers whose fathers had low education, and earnings increased by even more, 4.5%, for those whose fathers had low education but who themselves were of high ability. Conversely, the earnings of workers whose fathers were highly educated, and who were educated in the new system, fell by 5.6%. All of these sub-group effects were highly significant.

The key question for us is the degree to which the end of tracking, as opposed to the increase in compulsory schooling and the associated subsidy, could be responsible for any of these changes in earnings. The authors cite another paper which estimates that the return to one year of schooling in Sweden is 4.6%. If the observed earnings increases were all due to the observed increase in educational attainment, it would imply an 8.4% return to a year of education. One interpretation, if one accepts the outside estimate of the returns to a year of Swedish schooling, is that the reform package increased the returns to a year of schooling by 3.8%, perhaps by increasing school quality. But again, we cannot state whether any such increase in the returns to education emanates from the abolition of tracking rather than the other elements of the reform.

The finding that earnings of those whose fathers were highly educated fell is intriguing. One explanation might be that the increased heterogeneity of classrooms after the reform hurt the children of more highly educated parents, perhaps due to a weaker peer group for these children. But because the reform was implemented for entire municipalities at a time, we cannot rule out general equilibrium effects (namely, a drop in earnings induced by the increased supply of more educated workers). And as always in diff-in-diff models, a concern is that unobserved factors might have caused deviations in earnings between those who grew up in the treated areas and those who grew up elsewhere.

The United Kingdom acted in the 1960's to remove its system of tracking or "streaming" students into one of three levels of secondary education based on tests all students took at the age of eleven. On the surface this policy reform is cleaner than the Swedish reform because compulsory attendance laws did not change in the U.K. case. Still, some complications remained.

Numerous authors have studied the reforms in the U.K. Galindo-Rueda and Vignoles (2007) study a single cohort of students born in 1958, and thus do not implement a diff-in-diff estimator.

Rather, their approach is to model test scores and years of schooling completed at age 16 as functions of control variables observed at ages 11 and 7, including test scores. The key explanatory variable is whether a student attended a non- comprehensive (selective) school, and in alternative specifications the number of years the student was in a selective school. Because some Local Education Authorities (LEA's) allowed comprehensive and selective (grammar) schools to co-exist, the authors instrument the selective school indicator with the proportion of schools in the LEA that were comprehensive. Throughout, they use a propensity score matching approach in which they find that the political affiliation of the child's constituency is highly predictive of adoption of comprehensive schools. (Constituencies that elected a Conservative tended to be much slower to switch from grammar schools to comprehensive schools.)

Galindo-Rueda and Vignoles (2007) find that in their propensity score models, it matters whether they instrument the key explanatory variables related to attending selective schools. Without instruments, the results suggest positive effects of attending a selective school, and that students in the middle of the distribution gain the most. But the IV estimates, although supportive, are not statistically significant.

One of the more interesting findings of the paper is that including controls for achievement at age 11, right before students would enter the tracked system, greatly reduces the estimated effects of subsequent tracking. A natural interpretation is that selectivity bias has not been fully removed by the use of propensity scores. The authors develop a different interpretation, that in LEA's where selective grammar schools persisted, students had a stronger incentive to work hard before the age of 11 in the hope of entering a selective school at age 11. They buttress their theory by showing that LEA's that moved to a comprehensive system latest were the ones in which student achievement gains between ages 7 and 11 were higher.

Pischke and Manning (2006) re-analyze the data, and focus on the relative explanatory power of secondary school tracking on achievement gains between ages 11 and 16 (when the students were in secondary school) and the ages of 7 and 11. They conclude that because the "effect" of secondary tracking is of similar magnitude in these two age ranges, the elementary school effect cannot be real. Thus, they argue, selectivity bias must remain, even after they instrument for the date at which an LEA switches to comprehensive schools using political control of the county. However, it is conceivable that the gains in primary school could appear as large as the gains in secondary school, especially given that the tests at the different age groups are not vertically scaled.

It is not clear which story – incentive effects of secondary school tracking on primary school students (and their teachers), or selectivity bias – is the more important factor.

Pekkarinen, Uusitalo and Pekkala (2006) ask whether a move towards de-tracking in Finland in the 1970's is associated with changes in intergenerational income mobility. Their natural experiment appears to be very "clean". They examine a national school reform program implemented in phases between 1972 and 1977. In the pre-existing system, students were placed into one of two tracks after four years of primary school. After the reform, the "civic schools" that had enrolled many students up

to grade 8 or 9, and which provided a relatively vocational education, were abolished, as were most of the private secondary schools that enrolled students with strong academic aspirations. In their place, a nine-year comprehensive school for all students was implemented. (As before, after grade 9 students had an option to apply to upper secondary schools (the college-bound track) or vocational schools.) The reform was phased in over 5 years, with 6 broad regions being put on different timetables for reform.

Notably, by estimating models of sons' log earnings as functions of fathers' log earnings, the authors use a longer term outcome than test scores. They add interactions with sets of dummy variables for cohort and region and with the crucial dummy variable indicating whether the region had already implemented de-tracking by the time the son was in the age range affected by the reform. These dummy variables enter the equation directly and interacted with fathers' log earnings. The specific regressors are different from the set-up we provided in our sample data generation process in (2) but the identifying assumption is essentially the same: there cannot have been variations in the correlation between fathers' and sons' earnings by region that varied over time differently across groups.

They find that the introduction of comprehensive schools from grades 5 through 9 is associated with a 20% reduction in the relation between sons' and fathers' earnings, implying a strong increase in income mobility that resulted from de-tracking.

One final note on all of the above papers is that they all use geographic variations *within a country* to identify the effects of tracking. To the extent that families endogenously choose where to live within the given country, any effects of tracking may simply reflect endogenous sorting.

Brunello and Checchi (2007) take a similar difference in differences approach, using geographical and time variation in tracking policies, but unlike the above papers they compare different countries. They focus on whether tracking accentuates the relation between family background and long-term education outcomes such as educational attainment, postsecondary enrollment, literacy, and labor market outcomes including employment, training and wages.

The authors' decision to examine longer-term outcomes than test scores makes this paper an important contribution. One concern in the many papers that examine patterns of inequality in test score is that the way in which test scores are scaled could easily create the appearance of increased inequality. With longer-term measures like earnings and educational attainment, we have cardinal measures for which inequality measures are more naturally defined.

Their model in fact goes a step beyond our d.g.p. (2) by including dummies for country interacted with cohort. Their specification, because it does not measure grade-by-grade performance, does not include a subscript for grade g. To facilitate comparisons with (2) we use time t as an indicator for birth cohort:

(3) $$S_{ict} = \beta_{ct} + X_{ict}\Phi + F_{ict}\Pi + Q_{ict}\gamma + F_{ict}Q_{ict}\Omega + T_{ict}F_{ict}\rho + \left(\alpha_i + \varepsilon_{icgt}\right)$$

The random effect related to country and time, β_{ct}, is removed from the error term of (2) and specifically included as a fixed effect, which of course also removes the random effects related to country and those related to time. Similarly, we have dropped the error terms related to grade given that the authors focus on outcomes measured, roughly speaking, at the end of school (age 16) or outcomes measured for young adults. For the same reason, the tracking dummy T_{ict} in (3) is intended to capture the multiple-year effects expressed in (2).

Another innovation in (3) is that interactions between family background and a set of variables that are roughly analogous to measures of school quality Q_{ict} are included, to make sure that other variables are not affecting the slope of the family background variables in addition to tracking itself. [5]

The authors use two measures of tracking, "tracking length", that is, the number of years over which a student is likely to have been exposed to tracking, and the share of students in upper secondary schooling who are in vocational tracks. The first of these matches closely the tracking measures used by most other papers, while the latter variable is quite different, and merits scrutiny. A major concern is that the proportion of students in vocational education represents an endogenous outcome, the result of the interaction of the supply of student openings and demand for vocational schools, rather than a specific policy that could be argued to be exogenous.

Another concern is that the vocational education measure is prone to measurement error. For instance, Brunello and Checchi (2007) report that 0.0% of American high school students were in vocational education in 2002, (see their Table 1), on the grounds that virtually all students in the U.S. attend comprehensive schools. This misses the fact that tracking, including a robust vocational track, is a hallmark of American high schools. Levesque et al. (2008, Tables 2.1 and 2.2) report that although it is true that in the United States only 5.2% of high schools have a full Career and Technical Education focus, 88.1% of public high schools offer one or more occupational programs.

To be fair, this variable is not measured with more error than the age at which students are first tracked, which is widely used throughout the literature.

Mainly because of our concern that the percent of students in vocational tracks is endogenous, in our summary of the results of Brunello and Checci (2007) below we will focus on the "length of tracking" variable.

Brunello and Checchi (2007) find that for the most part, when countries track at an early age, the influence of parental education on long-term outcomes is accentuated. This applies to measures of educational attainment (years of schooling and the probability of dropping out), and to earnings and the probability of employment. Interestingly, they find that tracking reduces the influence of parental education on adult literacy and the probability of receiving on-the-job training. (Their finding on on-the-

[5] Brunello and Checchi in fact include some variables that we can directly interpret as measures of public school quality, such as the pupil-teacher ratio and expenditures on public education, as well as other variables measuring complementary or competing types of education: enrollment in private schools and enrollment in preschools.

job training is not necessarily counterintuitive – if less affluent students tend to take the vocational track, this may put them in line to receive more technical training once they launch their careers in jobs related to their upper secondary training.) Typically however, the interaction of tracking and parental education is only sometimes statistically significant. [6]

Geographic:Age Differences in Differences Using International and Sub-National Geographic Variation

Most international analyses of test scores and tracking have not used the Brunello and Checchi (2007) approach of exploiting changes within countries in tracking policies. This probably reflects the fact that that there has been little variation in tracking policies over the short time spans over which international tests of achievement have been available.

A common approach to this lack of temporal variation has been to replace the time element in a standard diff-in-diff formulation with a grade element. That is, the difference in test score gains between a grade that is untracked and a grade that is tracked, between countries with and without tracking, could plausibly be used as an estimate of the effect of tracking on test scores (or on the effect of tracking on the link between family background and achievement).

Most of the papers in this genre ask whether countries that track at an early age have higher transmission of inequality intergenerationally. Several reasons for such a relationship seem evident. Perhaps parents have quite a lot of influence over track placement when their children are in the lower grades, but less influence in the higher grades because a school will have a much better grasp on a student's true achievement after several years. Alternatively suppose that more educated parents tend to have more success in placing their children in high tracks, regardless of grade. [7] Suppose further that educational outcomes diverge between students in high and low tracks, and that this divergence increases with the length of time that students are separated into different tracks.

This geographical identification approach, much like that in the previous section, provides a deft solution to concerns about the endogeneity of tracking policies at the school level and the endogenous placement of students into ability groups. But concerns remain about the reasons for why countries or regions differ in their tracking policies. Omitted variable bias could account for any correlation across regions in student achievement and the use of tracking. Obviously, the potential for omitted variable bias is large in estimating such a model due to unobserved differences across disparate countries over time and grades.

[6] These results do not seem to hold when they instead model the inequality in earnings and literacy as a function of the dispersion within a country in family background. The former approaches seem more compelling because they use person-level observations.

[7] See Dustmann (2004) who documents a strong positive relation between parental education and children's secondary school track in Germany.

Hanushek and Woessmann (2006) provide a canonical example of this approach. They recognize that in their sample of countries, tracking never begins in primary grades and so there is some grade g^{min} below which tracking is never used. They use average test scores at the country level, for different grades, before and after tracking typically begins. The researchers have test scores available for a number of countries in grade g^H in year t and grade g^L in the year t' (which may be the same as t or in some instances earlier than t). They choose these grades such that in all countries $g^H > g^{min} > g^L$, so that achievement in g^L can be thought of as achievement in a grade before which any tracking has occurred. Then the average of these test scores, given the d.g.p expressed in (2), are:

$$\begin{aligned}\bar{S}_{cg^Ht} = {} & \bar{X}_{cg^Ht}\Phi + \bar{F}_{cg^Ht}\Pi + \bar{Q}_{cg^Ht}\gamma + T_{cg^Ht}\lambda_{g^H} + T_{c,g^H-1,t-1}\lambda_{g^H-1} + \mathrm{L} + T_{c,g^{\min_c},t-g^{\min_c}+1}\lambda_{g^{\min_c}} + \\ & T_{cg^Ht}\bar{F}_{cg^Ht}\rho_{g^H} + T_{c,g^H-1,t-1}\bar{F}_{c,g^H-1,t-1}\rho_{g^H-1} + \mathrm{L} + T_{c,g^{\min_c},t-g^{\min_c}+1}\bar{F}_{c,1,t-g^{\min_c}+1}\rho_{g^{\min_c}} + \\ & \left((\bar{\alpha}_i \mid g^H, t) + \alpha_c + \alpha_{g^H} + \alpha_t + \beta_{cg^H} + \beta_{ct} + \beta_{g^Ht} + \delta_{cg^Ht} + \bar{\varepsilon}^H_{cgt}\right)\end{aligned} \tag{4}$$

where implicitly the tracking indicators equal one for all of the grades listed, and

$$\begin{aligned}\bar{S}_{cg^Lt'} = {} & \bar{X}_{cg^Lt'}\Phi + \bar{F}_{cg^Lt'}\Pi + \bar{Q}_{cg^Lt'}\gamma \\ & + \left((\bar{\alpha}_i \mid g^L, t') + \alpha_c + \alpha_{g^L} + \alpha_{t'} + \beta_{cg^L} + \beta_{ct'} + \beta_{g^Lt'} + \delta_{cg^Lt'} + \bar{\varepsilon}^L_{cgt'}\right)\end{aligned} \tag{5}$$

Note that the latter equation does not include any tracking terms because g^L is a grade at which no country c has started tracking. Note also that in (4) tracking terms are added all the way down to the minimum grade at which country c begins to track, $g^{\min_c}$, and that this minimum grade varies from one country to another.

Hanushek and Woessmann (2006) model average test scores in one grade and year on average test scores in an earlier grade and nearby year, plus an indicator for early tracking ET_c. Let the regression coefficient on the test score in the earlier grade be denoted by b. By adding and subtracting b times the average score for grade g^L to the right hand side of (4), we gain a sense of what this regressor removes from the error term, and what biases likely remain:

$$\begin{aligned}\bar{S}_{cg^Ht} = {} & b\bar{S}_{cg^Lt'} + ET_c\tau \\ & + \left\{\begin{array}{l} \left(\bar{X}_{cg^Ht} - b\bar{X}_{cg^Lt'}\right)\Phi + \left(\bar{F}_{cg^Ht} - b\bar{F}_{cg^Lt'}\right)\Pi + \left(\bar{Q}_{cg^Ht} - b\bar{Q}_{cg^Lt'}\right)\gamma - ET_c\tau \\ + T_{cg^Ht}\lambda_{g^H} + T_{c,g^H-1,t-1}\lambda_{g^H-1} + \mathrm{L} + T_{c,g^{\min_c},t-g^{\min_c}+1}\lambda_{g^{\min_c}} + \\ T_{cg^Ht}\bar{F}_{cg^Ht}\rho_{g^H} + T_{c,g^H-1,t-1}\bar{F}_{c,g^H-1,t-1}\rho_{g^H-1} + \mathrm{L} + T_{c,g^{\min_c},t-g^{\min_c}+1}\bar{F}_{c,1,t-g^{\min_c}+1}\rho_{\min_c} \\ + \left[\begin{array}{l} \left((\bar{\alpha}_i \mid g^h, t) - b\left((\bar{\alpha}_i \mid g^L, t')\right)\right) + \alpha_c(1-b) + \left(\alpha_{g^H} - b\alpha_{g^L}\right) + \left(\alpha_t - b\alpha_{t'}\right) \\ + \left(\beta_{cg^H} - b\beta_{cg^L}\right) + \left(\beta_{ct} - b\beta_{ct'}\right) + \left(\beta_{g^Ht} - b\beta_{g^Lt'}\right) \\ + \left(\delta_{cg^Ht} - b\delta_{cg^Lt'}\right) + \left(\bar{\varepsilon}_{cg^Ht} - b\bar{\varepsilon}_{cg^Lt'}\right) \end{array}\right] \end{array}\right\}\end{aligned} \tag{6}$$

The early tracking indicator is included in the hope of picking up the average direct effects of tracking, that is, the λ terms, and the average family:tracking effects, that is, the ρ terms. Because different countries begin tracking in different grades $g^{\min_c}$, the regression coefficient τ should be interpreted as a weighted average of the effects of various years of tracking, which vary by country.

Above, the various terms inside the braces constitute the error term. Most of the terms in the braces will have non-zero expectation, and it seems likely that many of these components of the error term will be correlated with the early tracking indicator, thus biasing the coefficient estimate τ.

With roughly 16 country-pair observations, linear regression will of course choose b and τ to minimize the sum of squared residuals. Intuitively, if b=1, then most of the components in the error term reduce to differences in the mean values of, for example, average personal characteristics, $\left(\bar{X}_{cg^H t} - \bar{X}_{cg^L t'}\right)$, and the country fixed effect α_c would be completely removed. In practice, this coefficient is unlikely to be 1. Therefore this approach is not quite the same as adding a country fixed effect to the model, and remains, at least to some degree, prone to bias due to omitted country characteristics α_c. The omitted country traits are likely to be correlated with the use of tracking.

Most importantly, it seems highly likely that differences in school resources across grades and years could be linked to tracking. That is, ET_c could be correlated with $\left(\bar{Q}_{cg^H t} - b\bar{Q}_{cg^L t'}\right)\gamma$. In particular, if countries that track early also tend to increase/decrease spending per pupil more between primary and secondary school, and if spending per pupil is positively related to test scores, then the coefficient on early tracking could be biased upward/downward. It is not easy to control for the school spending factor exactly: recall that when we average our d.g.p, (2), across students, as in (4) and (5), the $\bar{Q}_{cgt}$ terms refer to cumulative school resources enjoyed by the average student between kindergarten and grade g. As mentioned earlier, Brunello and Checchi (2007, page 795) document large differences in pupil-teacher ratios for students in different tracks in various European countries, with academic students implicitly attending smaller classes than do vocational students. This does not establish that countries that track early tend to increase or decrease average resources disproportionately in higher grades, but it suggests that this is a strong possibility. The above concern applies not just to patterns of spending per pupil but to any other school characteristic that might change across grades (and/or tracks) such as teacher qualifications or class size.

Differences in the country/grade-specific error terms, $\left(\beta_{cg^H} - b\beta_{cg^L}\right)$, which in part will reflect the alignment between the test instrument and the curriculum taught in a given grade and country, could be strongly correlated with tracking. The difference in the country-year error terms, $\left(\beta_{ct} - b\beta_{ct'}\right)$, could reflect gradual changes in the average curriculum between years t' and t, which could be correlated, but not causally, with early tracking. (This problem is smaller if t is close to t', which is the choice Hanushek and Woessmann make.) Finally, the term $\left(\delta_{cg^H t} - b\delta_{cg^L t'}\right)$ could reflect unobserved causes of changes in achievement growth across grades within a country, for instance variations in the

preparation of teachers across grades over time, which we are liable to attribute to the existence of tracking.

Another approach to the problem of omitted variable bias is to use student-level data and to attempt to control for covariates. For example, Ammermüller (2005) studies reading achievement across countries in a model that examines the mediating influence of tracking on the effect of family background. He focuses on reading scores of students roughly in grade 9 and grade 4 across a set of countries, and like Hanushek and Woessmann (2007) exploits the fact that tracking does not begin until secondary school in his sample. His equation, using the above notation, translates roughly to:

$$S_{icgt} = \alpha_c + \alpha_g + X_{icgt}\Phi + Q_{icgt}\gamma + F_{icgt}\left(\Pi_c + \Pi_g\right) + T_{ic,9,t}F_{ic9t}\rho_9 + \left(\alpha_i + \alpha_t + \beta_{cg} + \beta_{ct} + \beta_{gt} + \delta_{cgt} + \varepsilon_{icgt}\right) \quad (7)$$

where the main tracking variable used, $T_{ic,9,t}$, equals the number of types of schools at the secondary level in country c at time t, for those in grade 9, and equals 0 for those in grade 4. This model incorporates a number of innovations designed to reduce bias. Most notably, the model incorporates both country and grade fixed effects, which are brought out of the error term in parentheses. However, a cost of this approach is that estimates from this model cannot be used to infer the overall effects of tracking on efficiency because the country fixed-effect removes any levels effect of tracking. [8]

Similarly, Waldinger (2006) adopts a student-level strategy, while distinguishing between grades before and after tracking has started in a country. He also distinguishes between an indicator for whether the country c uses early tracking, ET_c, and an indicator for whether the actual grade is one by which tracking has started. In his sample, he simply defines a binary variable SECONDARY which equals one for secondary grades, which makes sense because only in secondary grades has tracking started. In terms of the above notation, the estimating equation is approximately represented by:

$$S_{icgt} = \alpha_c + \alpha_g + X_{icgt}\Phi + F_{icgt}\left(\Pi + \Pi_g + ET_c\eta + T_{icgt}\rho_g\right) + Q_{icgt}\gamma + \left(\alpha_i + \alpha_t + \beta_{cg} + \beta_{ct} + \beta_{gt} + \delta_{cgt} + \varepsilon_{icgt}\right) \quad (8)$$

This approach is similar to that of Ammermüller (2005) in that it removes country and grade fixed effects, and allows the family characteristics to have separate effects by grade (but not by country). Although this model does not go as far as that of Ammermüller (2005) in fully interacting family background with country fixed effects, the effect of family background on test scores is allowed to shift by an identical amount η for all countries that track at an early age. The coefficient of interest is

[8] There are two other notable features about this equation. First it allows for the effects of family background to vary separately by grade and by country, which we have signaled above by including two separate vectors of coefficients Π. Second, the controls for school resources are quite limited, including variables like instructional time but no overall measure of spending per pupil or pupil-teacher ratios. This increases the possibility of omitted variable bias.

ρ_g which captures the effects of tracking specifically in grade g. (Because there are only two grades in the model, there is a single parameter estimated for the differential effect of being in a secondary school in a country that begins to track in secondary school.)

Findings on Inequality

Most of these papers support the hypothesis that family background is more strongly related to student outcomes in countries that track students at an early age. Thus, tracking generates inequality. For instance, Hanushek and Woessmann (2007) find that in models of various measures of test-score inequality, early tracking is associated with significant increases in inequality in secondary school relative to primary school. Ammermüller (2005) finds that not all interactions between measures of family socioeconomic status and the number of school types are statistically significant, but in roughly half the cases there was a significant relation supporting the idea that tracking increases the effect of family background. Most relevant was his uniform finding that parental education interacted with the number of school types was positively related to reading achievement in the higher (tracked) grades relative to the lower grades.

Waldinger (2006) finds an insignificant relation between parental education and whether the student was in a grade/country with tracking. (In equation (8) the null that $\rho_g=0$ is handily retained.) This finding is the opposite of the other studies noted above. However, η, the interaction between parental education and the use of early tracking, is routinely positive and highly significant. Similar results obtain when the background measure is based on books in the student's home.

There are at least three ways to interpret Waldinger's findings that $\rho_g=0$ and $\eta>0$. His own interpretation is that other studies that compare outcomes across countries failed to account for unobserved differences in the impact of family background across countries that do and do not track. Once one does take this into account by interacting family background with an indicator for whether the country as a whole tracks, there is no real increase in the importance of family background between untracked lower grades and tracked higher grades ($\rho_g=0$). This explanation could be correct. But a remaining concern is that (8) is in some ways a restricted version of Ammermüller's model (7) which allows for interactions between family background and a full set of country fixed effects. Yet this latter model, which should do an even better job of removing heterogeneity across countries in the effect of family background, finds that tracking does matter.

The two other explanations hearken back to the debate between Galindo-Rueda and Vignoles (2007) and Pischke and Manning (2006) in the context of the U.K. The conclusion of the former paper implies that perhaps the positive coefficients on the interaction between family background and whether the country as a whole tracks ($\eta>0$) indicates that during their primary school careers, students are selected and groomed for the type of school track they will enter in their secondary school years. In this case identification from a difference in difference based on country-grade contrasts may not be a valid approach. Alternatively, the Pischke and Manning (2006) paper implies that perhaps the positive

coefficient on η is another sign that fixed effect models cannot fully control for self-selection. The only difference is that Pischke and Manning (2006) focused on the selective nature with which jurisdictions in the U.K. switched to comprehensive schools; in the international setting the concern is that countries self-select into early tracking.

Findings on Efficiency

The only international study that directly assesses the impact of tracking on school efficiency is the paper by Hanushek and Woessmann (2006). (The others, by adding a fixed effect for each country, cannot estimate the direct effect of a tracking policy on test scores because of lack of any variation within country of the tracking policy during the time frames they study.) Hanushek and Woessmann (2006) cannot reject the null of not overall effect on achievement.

Other Methods that Use Geographical Variation in Tracking

Not all studies that exploit geographical variation fit into the two genres of difference in difference models discussed in the previous two sections.

A recent example is the work by Bauer and Riphahn (2006), which studies the correlation between parents' and children's education across Swiss cantons. (The cantons set the grade at which tracking begins.) The paper uses several measures of when tracking starts, to test for whether early tracking accentuates the intergenerational correlation between students' secondary school track and their parents level of education. The approach is cross-sectional, examining secondary school students from the 2000 census.

In terms of the d.g.p. outlined in (2), this approach virtually removes the random errors related to grade level α_g from the error term because all of the students are aged 17, and of course time variation α_t is removed as well because the data-set is a cross-section. But all of the components in the error term that relate to individuals i and region c remain. The main contribution that this approach makes to identification is that tracking is not measured at the level of the individual student or school, thus reducing concerns about endogeneity of tracking or students' individual tracks. But concerns remain about the potentially endogenous decision at the canton level of the grade at which to begin tracking, and the decision by families of the canton in which to reside. Similarly, the authors' model does not include any controls for school characteristics or demographics at the school or canton level. Given that a table in their paper shows that cantons that track at a later age tend to have more highly educated adult populations that spend a smaller share of public spending on education, omitted variable bias related to school resources and the geographical location (Q_{icgt} and α_c in (2)) could affect the results.

The paper reports that tracking that starts at a later grade is associated with increased probability that children enroll in the secondary track that is considered "university-bound", regardless of parental education level. (The authors do not claim that this is a sign that late-tracking increases efficiency. Such a conclusion would be unwarranted because the models do not control for school resources.) Second, on the question of inequality, the results suggest that starting to track at a later age reduces relative differences by parental education in the probability that the student is in the highest secondary school track. Thus, again a paper that geographically aggregates the tracking variable again finds that tracking aggravates inequality.

Woessmann (2007) uses a cross- section of 16 observations from German states and finds a negative relationship between the use of late tracking and the slope of the (positive) gradient between math test scores and the socioeconomic status of the student.

Schutz, Ursprung and Woessmann (2008) compile an unusually large set of 54 nations and test whether the link between students' math scores and books in the home, which is strongly positive, becomes weaker in countries that begin tracking at a later age. They find evidence favoring this idea. Although this paper does not use either of the difference in difference strategies typical of many of the

international papers, and so risks omitted variable bias, it has roughly two to four times as many countries as in many of the earlier studies. Another worthwhile feature of this paper is that it explicitly controls for immigration in its estimate of the slope between test scores and family background.

Endogeneity of Tracking in International Studies

Further work might gainfully test for endogeneity of tracking in international studies, especially the approaches that do not use a time difference that exploits changes in tracking policy within countries.

One reasonably plausible scenario is that some social or political factor, unmeasured by the researcher, influences both the age at which tracking starts and the gradient of achievement with respect to family background (dS/dF) in the opposite direction. This could induce a spurious negative correlation between the two.

To make this issue concrete, consider immigration. Suppose that researchers used parental education as their measure of family background F, and calculated dS/dF without taking into account whether the parents were immigrants. Then countries with large numbers of immigrants might be observed to have higher dS/dF gradients, because unknown to the researcher parents with low education are likely to also not speak the native language fluently or know how best to negotiate public schools on behalf of their children. Suppressing the subscripts from the earlier models, we can express achievement S in a country as a function of family background F, and the average immigrant-to-population ratio I, where all of the a_j coefficients in the following model are positive:

(9) $$S = a_1 + a_2 F + a_3 F * I - a_4 I$$

Thus $dF/dS = a_2 + a_3 I$ is an increasing function of a nation's immigrant-to-population ratio.

Consider next the evidence in Betts and Fairlie (2003) that in metropolitan areas in the United States, increases in the immigrant-to-population ratio among young people is associated with increases in the probability that natives enroll their children in private schools. If this relationship were causal, it is easy to imagine how native parents might also react to inflows of immigrants by seeking greater (or earlier) within-school ability grouping. Thus if T is our measure of age at which tracking starts, $dT/dI < 0$.

Putting this endogenous relationship of tracking to immigration together with the positive interaction between immigrant share and dS/dF, it becomes possible that models that don't take these relations into account would detect a non-causal negative relation between the age at which tracking begins and the achievement:family background gradient.

Notably, Schutz, Ursprung and Woessmann (2008) take into account the immigration status of parents of individual students and allow for an interaction with their main measure of family background, and still find that countries that begin to track at a later grade have less inequality in

achievement. The only loose thread in this particular example would be to study whether the age at which tracking starts immigrant shares in the population is an endogenous function of immigration.

Another of the many possible socioeconomic or political factors that could influence both tracking and the achievement:background gradient is political beliefs. Galindo-Rueda and Vignoles (2007) and Pischke and Manning (2006), in the context of the U.K., and Figlio and Page (2002), in the context of the United States, present evidence that a local electorate that votes more conservatively is associated with use of tracking at an earlier age. Suppose that this pattern obtained internationally. Perhaps more conservative societies tend to prefer other policies that tend to strengthen the link between achievement and family background, for instance through policies related to housing segregation, or public subsidies for pre-school. If these policies are not perfectly measured by the researcher, then even if there is no causal relation between tracking and the gradient between achievement and family background, the regression model is likely to suggest a positive relation between use of tracking and this gradient.

These examples are speculative, but are intended to illustrate the many ways that endogeneity of tracking, combined with omitted variable bias, could lead to incorrect inference. Much could be done to search for factors that explain variations across countries in tracking policies, and to explore, as in Schutz, Ursprung and Woessmann (2008), what factors mediate the relationship between student outcomes and family background. Also, to the degree that omitted demographic and political variables are unchanging over time, international studies that exploit changes in tracking over time might reduce the potential for such problems.

Experiments with Random Assignment of Students to Treatment

As discussed above, comparisons across countries lead to concerns about major unmeasured differences that could bias results. Yet most existing within-country studies raise questions about why one school uses tracking and another does not, and whether unobserved characteristics of students are correlated with track placement. An additional concern in all of these studies is whether the definition of tracking is the same across schools or countries.

An apparent antidote to these problems would be to design an experiment in which schools were randomly assigned to tracked or non-tracked status, and in which a consistent rule was used to assign students to classrooms in schools that tracked. Further, one would want to ensure that schools in the control group did not stealthily implement their own form of tracking, so that one could prevent substitution bias. [9] A properly executed experiment removes biases stemming from the correlation between tracking and the error terms in (2) that arises in non-experimental settings. It does this by randomly assigning treatment status, rendering the tracking variables in (2) exogenous.

[9] See section 5.2 of Heckman, LaLonde and Smith (1999) for an overview of substitution bias in the context of evaluating government training programs.

Slavin (1987, 1990) reports that a total of 15 such experiments have been conducted in American secondary schools and that one experiment has been conducted in an American elementary school. These experiments were conducted from the 1920's through the early 1970's, after which time Slavin could find no new U.S. experiments.

More recently, a large scale experiment has been conducted in Kenya. We begin with that experiment and then discuss the 16 experiments conducted in the United States.

A recent paper by Duflo, Dupas and Kremer (2008) study an experiment in Kenya. World Bank funding allowed 121 elementary schools that had a single grade 1 class to receive an additional grade 1 teacher for an 18-month period. (In the second year of the program, the same cohort of students continued to participate in the program, and the additional teacher moved with these students to grade 2.) In one half of schools, a pre-test administered by the local schools was used to rank students, and students were assigned in strict order to the high-ability and the low-ability classrooms. To create a control group against which to compare the students in the tracked schools, in the other half of schools students were randomly assigned to the two classrooms. As the authors point out, this second procedure guards against the concern expressed by Betts and Shkolnik (2000b) that in non-experimental settings supposedly untracked classrooms and schools really are using ability grouping in some disguised form.

Tests in math and literacy were administered to the students 18 months after the arrival of the extra teachers. Also, in order to test whether any effects of tracking persisted, the students were tested a second time, one year after the program had ended.

The authors perform the standard tests to ensure that the control and treatment samples were similar, and that no students switched between classrooms.

The paper addresses several distinct questions. First, the authors examine whether the overall test scores in the schools with tracking are higher or lower after 18 months, which can be interpreted as a quite clean test of whether tracking has any impact on the efficiency of schools. Second, they examine whether tracking affects the distribution of test scores. Third, they use two different designs to study the impact of small and large variations in classroom peers. The chapter by Sacerdote in this volume covers the latter questions thoroughly, and we will instead focus mostly on the authors' analysis of the overall effect of tracking. Nonetheless, the authors' experimental design generates some compelling evidence about possible mechanisms through which tracking manifests its effects.

On the question of efficiency, the paper reports statistically significant increases in average test scores in the tracked schools of 0.175 standard deviations relative to the untracked schools, after controlling for covariates. Second, these effects largely persist a full year after the end of the program.

Third, on the question of tracking and inequality, the results stand in stark contrast to most of the inter-country literature. All ability groups, when defined by quartiles of initial test scores within schools, seem to gain equally from tracking. However, the coefficients on the ability/tracking interactions, while not nearing statistical significance, do hint that students in the higher quartiles may

gain slightly more than students in the lower quartiles. Due to the lack of precision of these interactions, it is not clear whether the proper interpretation is that tracking benefited all groups equally or that tracking increased inequality slightly across student groups.

The paper also exploits the experimental design to infer the mechanisms through which tracking was helping students. The authors find evidence that small changes in peer groups affect achievement. Thus, some other factor must more than counterbalance the negative peer effect that students placed in the low-ability classes experience. The paper reports some evidence that teachers were able to take advantage of the smaller variance within their classrooms by focusing on the skills that students have yet to learn. Evidence backing this idea is that students in the lower ability tracked classroom show greater growth in simpler math concepts than their counterparts in untracked schools, while students in the higher ability classrooms improve relatively more in more advanced concepts. The implication is clear: lower-ability students appear to have gained from tracking – even though their peer group had lower achievement than if the school had not used tracking – because teachers can do a better job when they teach a more homogeneous group of students.

The paper also finds that teachers in tracking schools are more likely to come to work and to be in the classroom, although this result is limited to the upper-ability track and to civil-service teachers who had relatively secure employment. Recall the argument in the theoretical section above that under the existence of tracking, all other school resources experienced by the individual student could adjust, including teacher effort. It appears that teacher effort may indeed adjust, at least in this setting.

A commonly expressed concern about experiments is that although they may have high internal validity, their external validity, that is, applicability to other situations, is not always high. It seems apparent that extrapolation to developed countries is problematic. The average class size was around 46 after the experiment began. One sign of the large heterogeneity within each grade was the standard deviation in age of about 1.5 years. Still, these conditions may mimic quite well school conditions in other developing countries.

To their credit, the authors also point out that randomly assigning teachers to classrooms may not mimic the real-life implementation of tracking.

A final concern about external validity is that by hiring contract teachers whose job security was far lower than the government teachers who typically teach in Kenyan schools, the experiment did not use a typical sort of teacher or a typical teacher contract, which raises questions about external validity not just for other countries but for Kenya itself.

In comparing these results to the large number of studies that make international comparisons, reviewed in the previous section, it is also important to remember that these latter studies examine the effects of tracking at the secondary level, not the effects of tracking in grades 1 and 2 as in the study by Duflo, Dupas and Kremer (2008). Not only is there an age difference, but the types of tracking typical in the secondary school systems in Europe involved not just ability grouping but the exposure of students to quite distinct curricula and schools.

In conducting an experiment on tracking at the elementary level, the Duflo, Dupas and Kremer (2008) paper seems to have only one precedent, even though there are many experiments that have been conducted in the U.S. at the secondary level. In that experiment, Cartwright and McIntosh (1972) report that elementary school students in Hawaii were randomly assigned to heterogeneous classes, to ability grouping across grades, and to "flexible" grouping, in which students were grouped separately by subject, again with the possibility of putting students in different grades into the same classroom if their achievement levels were similar. The authors find slightly negative overall effects of either approach to tracking. However, with only 262 students in one school, all of whom were disadvantaged, the study may not generalize to other settings.

Slavin (1990) reviews 15 tracking experiments conducted in the United States in secondary schools. In six of these experiments, students were randomly assigned to tracked or untracked classrooms, with assignment to ability group being decided by students' initial test scores or grades. In the other nine experiments, students were pre-matched based on achievement, and then randomly divided into tracked or untracked classrooms. On the question of the overall effect of tracking on achievement, Slavin finds that four of the studies showed positive effects, two showed no effect, and nine showed negative effects. In most cases, the estimated effects were quite small to moderate, with an absolute effect size of 0.01 to 0.3 standard deviations. One outlier, a study of 240 students in Virginia, produced an effect size of -0.48 standard deviations by the end of a one-year experiment. The largest positive effect found, in a study of the math achievement of 148 students in Ohio, reported an effect of 0.28 standard deviations, resulting from a single semester of tracking.

All but three of these experiments conducted in secondary schools tested for differential effects on students in the various ability groups. Six of the 12 studies suggest that students in high ability group gained the most and that students in the low ability group gained the least. However, four studies suggested that the high(est) ability group gained less than the other groups.

With two exceptions, each of the experiments lasted for one year. This of course, as in the Kenyan study, raises concerns that tracking could have a quite different effect if teachers believed that the tracking (or detracking, as the case may be) represented permanent reforms rather than ephemeral experiments.

One experiment which stands out both for the relatively large size of the student sample and the duration of the experiment is the work by Marascuilo and McSweeney (1972). In this experiment conducted in Berkeley, California, 603 students were randomly assigned to a high-medium-low system of ability grouping or to ungrouped social studies classes, and were tested with both standardized tests and tests created by the teachers involved in the experiment. The overall effect was -0.22 of a standard deviation, and evidence suggested that students in high-ability classes gained 0.14 of a standard deviation, while those in medium- and low-ability classes lost -0.37 and -0.43 of a standard deviation relative to their similar-ability counterparts in the control group. The same patterns obtained for the teacher-created test and the standardized test, although significance levels were lower in the latter case (5% versus 10% respectively).

Slavin (1990) cites an important limitation of the secondary school experiments: they did not allow schools to offer a different curriculum to students in the various ability groups. In the United States, secondary school tracking certainly does allow for curricular differences across grades, and so these experiments may not be at all representative of how tracking in U.S. secondary schools plays out. In particular, they may understate the differential effects on achievement.

Conclusion and Outline of a Possible Research Agenda for the Future

In spite of many decades of research, what we do not know about the effects of tracking on outcomes greatly exceeds what we do know. Our uncertainty reflects not only the usual methodological debates about causal inference, but also, and perhaps more fundamentally, quite poor measures of tracking combined with differences across countries in what tracking really means.

One always hopes to resolve any differences in findings by showing that they result largely from differences in method. Once one has ranked the methods, then the findings that emerge from the best method can be taken as the closest approximation to the truth. That approach does not apply readily in the case of tracking, because different methods have been used in different geographical contexts, and it is quite clear that definitions of tracking and ability grouping differ from one country to the next. Worse, existing measures of tracking do a rather poor job of identifying what these differences are, either between nations in the international studies, or between schools in national or local studies.

With these concerns in mind, what can we say about the various methods used to date and the empirical findings? Second, can we lay out some viable research paths for the future?

The largely American (and to a lesser extent, British) literature that examines tracking on a school-by-school basis offers the opportunity to control for confounding variables such as school resources, local area characteristics and school demography. But these advantages come at a high cost – the need to control for the endogeneity of whether schools track and the track into which students are placed. The early literature, as reviewed by Slavin (1987, 1990) shows little evidence that the use of tracking affects the overall effectiveness of schools, or changes inequality in student outcomes. A number of more recent papers in the U.S. contradict that view, suggesting that tracking materially aggravates inequality. However, more recent papers find that once one controls for endogenous track placement, these large effects either become much smaller or disappear altogether (Betts and Shkolnik, 2000a,b and Figlio and Page, 2002). Figlio and Page (2002) also present some evidence that treating the use of tracking as endogenous can even generate the result that the use of tracking lessens inequality.

The bulk of work recently has instead focused on tracking at a regional or national level. This approach solves the problem of endogenous placement of students into tracks, and lessens but does not eliminate the problem of endogenous tracking policies. These advantages come at a cost, namely, the potential for very large omitted variable biases. Roughly speaking there have been three ways in which regional or national variation in tracking has been exploited. On the surface the most convincing

approach has been to use changes in tracking policy over time, allowing a traditional differences-in-differences approach. Some of these studies have produced important hints that the decision to group students heterogeneously for longer periods of time leads to better outcomes for students from disadvantaged backgrounds. However, the two papers studying the transition to comprehensive secondary schools in the United Kingdom raise serious concerns that the date at which each jurisdiction adopted comprehensive schools is endogenous. Both find that even before the age at which tracking begins, in areas with tracking, young students progress more quickly than do students in other areas. Two logical explanations have been put forward – either the spectre of secondary school tracking induces primary school students to work harder, or the decisions on whether to track are endogenous, and existing methods fail to control for this endogeneity adequately.

The second large set of papers that uses regional variation in tracking does not use changes in tracking policy, but instead compares outcomes in grades before and after tracking begins. Because most of this research has added dummies for region, nothing can be concluded about the overall effect of tracking on achievement. However, one paper that uses a quasi difference in difference approach by regressing outcomes in high grades on outcomes in low grades suggests that the efficiency effects are near zero. What all of these papers can readily do is to address the inequality question. The papers fairly uniformly conclude that early tracking exacerbates differences in achievement that are correlated with family background. Here again, though, it may be that if countries that track at the secondary level, this induces changes in behavior in primary schools. Such incentive effects raise questions about the validity of using primary school students as controls for secondary school students within a given country.

Third, some papers use cross-sectional variation but do not use variation in tracking policy across time or grade. These are useful papers but are particularly likely to suffer from omitted variable bias.

Another issue related to all of the international papers is that in some ways, even if student-level data are available, the true sample size is the number of countries, because this is the level at which tracking varies.

Finally, we have a large set of tracking experiments performed in the United States mostly before 1970, and one recent tracking experiment performed in Kenya. Taken as a whole, the former studies suggest no meaningful effects of tracking on overall achievement or on inequality in achievement. The recent Kenyan study suggests meaningful productivity gains from tracking, ostensibly due to the heightened effectiveness of teachers when they teach a more homogenous class, and no significant effects on inequality. All of the experiments raise concerns about external validity. The American experiments in secondary schools focused on changing peer groups, but did not change curriculum between treatment and control groups. The treatments typically were very short, about one year. The Kenyan experiment was a short-term experiment in which teachers were randomly assigned to classes, and the newly hired teachers lacked the job security of the other teachers. Moreover the population being studied, and average school resources, were starkly different from what one sees in European or North American schools.

To sum up, on the question of the overall efficiency effect of tracking, the school-level literature does not find big effects; only one of the international studies uses a specification that allows the efficiency question to be addressed, and finds no effect; experimental evidence from the U.S. suggests no large effect, and the Kenyan experiment does suggest a fairly large positive effect.

On the distributional question, some of the more recent school-level research suggests that tracking aggravates inequality. But the identification in these papers has been questioned, with later papers finding a smaller or insignificant effect. The early school-level literature suggests no effect. The international and regional studies suggest that tracking aggravates inequality, while the experimental literature suggests there is no distributional effect.

The fact that the American literature tends to find less evidence that tracking generates inequality than does the international literature may reflect differences in resources. Brunello and Checchi (2007) cite evidence that different types of secondary schools in some European countries receive quite different resource packages. In the United States, where tracking occurs mainly within schools, evidence generated by Betts and Shkolnik (2000a) and corroborated by Rees, Brewer and Argys (2000) point to quite small differences across ability groups in teacher qualifications and class size, and indeed, countervailing differences. Further, the European tradition of tracking often involves sending vocationally and academically tracked students to different schools, which further creates potential for differential outcomes. Put differently, the large effects on inequality that emerge from many of the international studies could be real, and could stand out more than in the American literature because of the relative lack of variation in school resources and curriculum across ability groups within America relative to that in some other countries.

In the end, definitional differences and poor measures of tracking likely have more to do with the lack of agreement in the literature than do differences in methodology. The American literature suggests that asking principals and teachers whether a school tracks can yield different answers. We also lack precise ways of asking teachers about the ability level of their classes. But perhaps the most glaring example of measurement issues is that in most international studies of tracking, American schools are listed as "late trackers" or, in some studies, are listed as started tracking at the university level. Yet the school-level U.S. literature finds evidence of widespread ability grouping and tracking, in the sense of students being grouped by achievement and taking different curricula, from grade 6 or the start of secondary school, forward. Canadian schools are similar in this regard. Thus, the papers that compare countries may in fact be comparing two distinct forms of tracking, across-school tracking, as is or was common in Europe, with within-school tracking, which occurs in some countries that have been mislabeled as "late trackers".

A good starting point for further work would be to focus on better measurement. Surveys and observation at the school level would give us a far richer understanding of how tracking varies across areas. For example we need better information on whether tracking is mere ability grouping within grades, or curricular tracking; grouping within versus across grades; overall grouping or separate grouping by subject; grouping within or between schools; grouping that leads to dramatic differences in

the teachers, class size and curriculum experienced by students in different groups, or grouping without much tailoring of school resources or curriculum.

Once we had this information we would be in a much better position to understand what facets of tracking are driving apparent differences in student outcomes across countries or regions within a country.

What about the methods themselves? Which should be the focus for the future?

Further work that measures tracking at the school level will not make important contributions unless it adopts methods to control for endogeneity of student placement and schools' decisions to track. Clearly, the measures of tracking need to improve. One rather expensive solution is to undertake detailed national studies involving surveys and classroom observation to document better the nuances of how tracking works in different locations. Another possible strategy is to avoid relying solely on principal or teacher surveys and instead to observe the classroom groupings of individual students as a function of various measures of their past academic achievement and courses completed. This approach has recently become more feasible thanks to the growing number of panel data-sets in the United States that follow all students within a district, with details on classroom assignments.

In the international literature, better information on what tracking really is, careful thought about why countries vary in tracking policy, and a greater emphasis on finding policy innovations over time all would help convince readers that something close to causal is being measured.

The experimental literature has by far the best claim that it is isolating the causal effects of treatment. But it is easy to question the relevance of these experiments to the real world. What if the treatment is not a realistic portrayal of how tracking is actually implemented? Slavin (1990) expresses an important point: the many experiments in American secondary schools contrasted only the peer groups between treatment and control groups, while not allowing for the fundamental differences in curriculum that Oakes (1985, 2000) documents in American schools. Further, if teachers know that an experiment will last for only a year, and if the experiment omits other important real-world considerations, the external validity of the experiment falls into question.

This is why we need detailed studies of how tracking works in the real world in all sorts of variants. Then one could experimentally evaluate all of these real-world flavors of tracking.

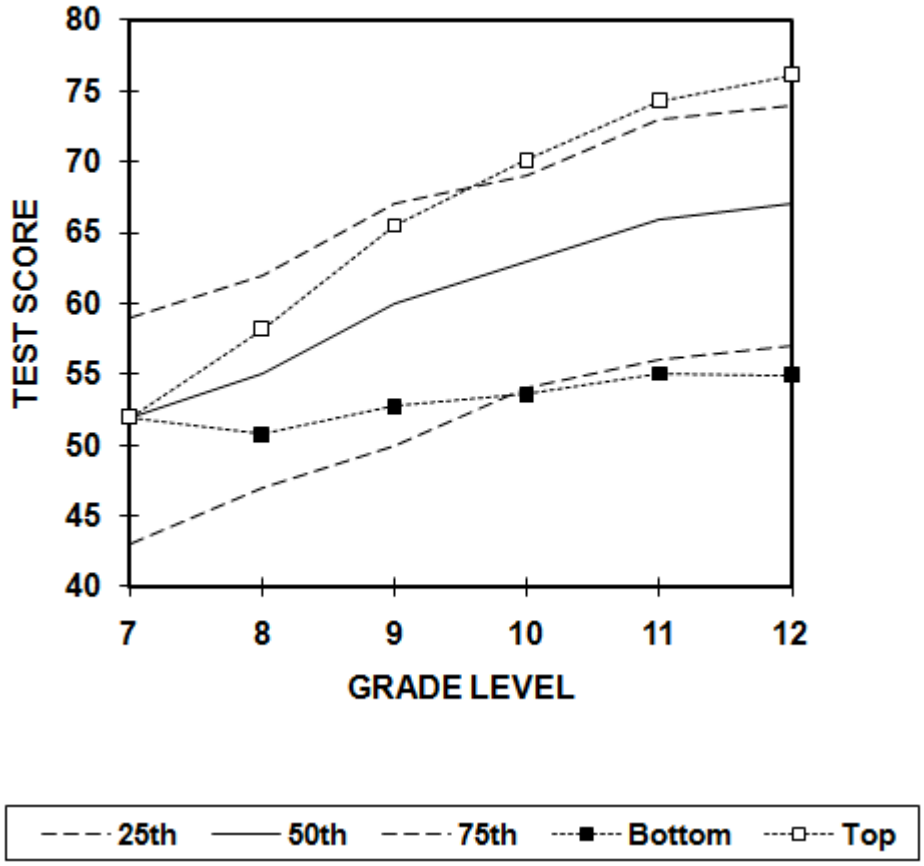
If researchers succeeded in opening up the black box by documenting the details on how different areas implement tracking, this could allow for a set of quite realistic experiments. In the United States, given how prevalent tracking already is, one approach might be to sample a large number of school districts that had been documented to use different flavors of tracking, to use the status quo as the treatments, and then to a random sample of schools to a control group in which students were instead grouped heterogeneously.

One could imagine a series of experiments, each within a given country, in which some of the treatments were new, perhaps more radical, and borrowed from practices in other countries. For

instance, European nations might experiment with more American-style within-school tracking, which could yield a test of whether the apparent effects on inequality of European-style streaming could be mitigated if streaming occurred within rather than across schools.

Ultimately, if the interventions were designed carefully, we might learn that in different forms, tracking can either increase or decrease efficiency, and can either increase or decrease inequality. The policy prescription would then emerge from our new, far more nuanced understanding of the many varieties of tracking. In the absence of such developments, continuing to treat tracking as a black box, which can be fully modeled by a single dummy variable, could well be a waste of researchers' time and funders' money.

Figure 1 Actual Range of Test Scores, and Predicted Scores of Identical Individuals Placed in Top and Bottom Classes



Source: Betts and Shkolnik (2000a).

References

Ammermüller, Andreas. "Educational Opportunities and the Role of Institutions." Centre for European Economic Research Discussion Paper No. 05-44 (2005).

Argys, Laura M., Daniel I Rees, and Dominic J. Brewer. "Detracking America's Schools: Equity at Zero Cost?" Journal of Policy Analysis And Management 15, no. 4 (1996): 623-45.

Babcock, Philip, and Julian R. Betts. "Reduced-Class Distinctions: Effort, Ability, and the Education Production Function." Journal of Urban Economics 65, no. 3 (2009): 314-22.

Bauer, Philipp, and Regina T Riphahn. "Timing of School Tracking as a Determinant of Intergenerational Transmission of Education." Economics Letters 91, no. 1 (2006): 90-97.

Betts, Julian R., and Robert W. Fairlie. "Does Immigration Induce 'Native Flight' from Public Schools into Private Schools?" Journal of Public Economics 87, no. 5-6 (2003): 987-1012.

Betts, Julian R., and Jamie L. Shkolnik. "The Behavioral Effects of Variations in Class Size: The Case of Math Teachers." Educational Evaluation and Policy Analysis 21, no. 2 (1999): 193-213.

———. "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools." Economics of Education Review 19, no. 1 (2000a): 1-15.

———. "Key Difficulties in Identifying the Effects of Ability Grouping on Student Achievement." Economics of Education Review 19, no. 1 (2000b): 21-26.

Betts, Julian R., Andrew Zau, and Lorien Rice. Determinants of Student Achievement: New Evidence from San Diego, San Francisco: Public Policy Institute of California, 2003. Available at www.ppic.org.

Brunello, Giorgio, and Daniele Checchi. "Does School Tracking Affect Equality of Opportunity? New International Evidence." Economic Policy 22, no. 52 (2007): 781-861.

Cartwright, G Phillip, and Dean K McIntosh. "Three Approaches to Grouping Procedures for the Education of Disadvantaged Primary School Children." Journal of Educational Research 65, no. 9 (1972): 425-29.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. "Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." NBER Working Paper 14475 (2008).

Dustmann, Christian. "Parental Background, Secondary School Track Choice, and Wages." Oxford Economic Papers 56, no. 2 (2004): 209-30.

Epple, Dennis, Elizabeth Newlon, and Richard Romano. "Ability Tracking, School Competition, and the Distribution of Economic Benefits." Journal of Public Economics 83 (2002): 1-48.

Figlio, David N., and Marianne E. Page. "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" Journal of Urban Economics 51, no. 3 (2002): 497-514.

Finn, Jermey D., and Charles M. Achilles. "Tennessee's Class Size Study: Findings, Implications, Misconceptions." Educational Evaluation and Policy Analysis 21, no. 2 (1999): 97-109.

Galindo-Rueda, Fernando, and Anna Vignoles. "The Heterogeneous Effect of Selection in Uk Secondary Schools." In Schools and the Equal Opportunity Problem, edited by Ludger Woessman and Paul E. Peterson, 103–28. Cambridge, MA: MIT Press, 2007.

Gamoran, Adam, and Robert D. Mare. "Secondary School Tracking and Educational Inequality: Compensation, Reinforcement, or Neutrality?" American Journal of Sociology 94, no. 5 (1989): 1146-83.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor. "Aggregation and the Estimated Effects of School Resources." Review of Economics and Statistics 78, no. 4 (1996): 611-27.

Hanushek, Eric A., and Ludger Woessmann. "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries." Economic Journal 116, no. 510 (2006): C63-C76.

Heckman, James J., Robert J. Lalonde, and Jeffery A. Smith. "The Economics and Econometrics of Active Labor Market Programs." In Handbook of Labor Economics, edited by Orley Ashenfelter and David Card, 1865-2097. Amsterdam: North Holland, 1999.

Hoffer, Thomas B. "Middle School Ability Grouping and Student Achievement in Science and Mathematics." Educational Evaluation and Policy Analysis 14, no. 3 (1992): 205-27.

Kerckhoff, Alan C. "Effects of Ability Grouping in British Secondary Schools." American Sociological Review 51, no. 6 (1986): 842-58.

Krueger, Alan B. "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics 114, no. 2 (1999): 497-532.

Levesque, Karen, Jennifer Laird, Elisabeth Hensley, Susan P. Choy, Emily Forrest Cataldi, and Lisa Hudson. Career and Technical Education in the United States 1990 to 2005. NCES 2008-035.Washington, D.C.: National Center for Education Statistics, U.S. Department of Education, 2008.

Marascuilo, Leonard A, and Maryellen McSweeney. "Tracking and Minority Student Attitudes and Performance." Urban Education 6 (1972): 303-19.

Meghir, Costas, and Marten Palme. "Educational Reform, Ability, and Family Background." American Economic Review 95, no. 1 (2005): 414-24.

Oakes, Jeannie. Keeping Track: How Schools Structure Inequality. New Haven: Yale University Press, 1985.

———. Keeping Track: How Schools Structure Inequality. Second ed. New Haven: Yale University Press, 2005.

Pekkarinen, Tuomas, Roope Uusitalo, and Sari Pekkala. "Education Policy and Intergenerational Income Mobility: Evidence from the Finnish Comprehensive School Reform." IZA Discussion Paper No. 2204, 2006.

Pischke, Jorn-Steffen, and Alan Manning. "Comprehensive Versus Selective Schooling in England in Wales: What Do We Know?" NBER Working Paper Series,no. 12176 (2006), http://www.nber.org/papers/w12176.

Rees, Daniel I., Dominic J Brewer, and Laura M. Argys. "How Should We Measure the Effect of Ability Grouping on Student Performance?" Economics of Education Review 19, no. 1 (2000): 17-20.

Schutz, Gabriela, Heinrich W. Ursprung, and Ludger Woessmann. "Education Policy and Equality of Opportunity." Kyklos 61, no. 2 (2008): 279-308.

Slavin, Robert E. "Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis." Review of Educational Research 57, no. 3 (1987): 293-336.

Slavin, Robert E. "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis." Review of Educational Research 60, no. 3 (1990): 471-99.

Theil, Henri. Linear Aggregation of Economic Relations, Contributions to Economic Analysis, 7. Amsterdam,: North-Holland Pub. Co., 1954.

Waldinger, Fabian. "Does Tracking Affect the Importance of Family Background on Students' Test Scores?" Unpublished manuscript, London School of Economics (2006).

Woessmann, Ludger. "Fundamental Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries." CESifo Working Paper 1981, (2007).